Expert Systems With Applications 44 (2016) 156–167

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Mathematical programming for piecewise linear regression analysis

Lingjian Yang [a], Songsong Liu [a], Sophia Tsoka [b], Lazaros G. Papageorgiou [a,*]

[a] Centre for Process Systems Engineering, Department of Chemical Engineering, UCL (University College London), Torrington Place, London WC1E 7JE, UK
[b] Department of Informatics, School of Natural and Mahtematical Sciences, King's College London, Strand, London WC2R 2LS, UK

ARTICLE INFO

Keywords:
Regression analysis
Surrogate model
Piecewise linear function
Mathematical programming
Optimisation

ABSTRACT

In data mining, regression analysis is a computational tool that predicts continuous output variables from a number of independent input variables, by approximating their complex inner relationship. A large number of methods have been successfully proposed, based on various methodologies, including linear regression, support vector regression, neural network, piece-wise regression, etc. In terms of piece-wise regression, the existing methods in literature are usually restricted to problems of very small scale, due to their inherent non-linear nature. In this work, a more efficient piece-wise linear regression method is introduced based on a novel integer linear programming formulation. The proposed method partitions one input variable into multiple mutually exclusive segments, and fits one multivariate linear regression function per segment to minimise the total absolute error. Assuming both the single partition feature and the number of regions are known, the mixed integer linear model is proposed to simultaneously determine the locations of multiple break-points and regression coefficients for each segment. Furthermore, an efficient heuristic procedure is presented to identify the key partition feature and final number of break-points. 7 real world problems covering several application domains have been used to demonstrate the efficiency of our proposed method. It is shown that our proposed piece-wise regression method can be solved to global optimality for datasets of thousands samples, which also consistently achieves higher prediction accuracy than a number of state-of-the-art regression methods. Another advantage of the proposed method is that the learned model can be conveniently expressed as a small number of if-then rules that are easily interpretable. Overall, this work proposes an efficient rule-based multivariate regression method based on piece-wise functions and achieves better prediction performance than state-of-the-arts approaches. This novel method can benefit expert systems in various applications by automatically acquiring knowledge from databases to improve the quality of knowledge base.

1. Introduction

In data mining, regression is a type of analysis that predicts continuous output/response variables from several independent input variables. Given a number of samples, each one of which is characterised by certain input and output variables, regression analysis aims to approximate their functional relationship. The estimated functional relationship can then be used to predict the value of output variable for new enquiry samples. Generally, regression analysis can be useful under two circumstances: (1) when the process of interest is a black-box, i.e. there is limited knowledge of the underlying mechanism of the system. In this case, regression analysis can accurately predict the output variables from the relevant input variables without requiring details of the however complicated inner mechanism (Bai, Wang, Li, Xie, & Wang, 2014; Cortez, Cerdeira, Almeida, Matos, & Reis, 2009; Davis & Ierapetritou, 2008; Venkatesh, Ravi, Prinzie, & den Poel, 2014). Quite frequently, the user would also like to gain some valuable insights into the true underlying functional relationship, which means the interpretability of a regression method is also of importance, (2) when the detailed simulation model relating input variables to output variables, usually via some other intermediate variables, is known, yet is too complex and expensive to be evaluated comprehensively in feasible computational time. In this case, regression analysis is capable of approximating the overall system behaviour with much simpler functions while preserving a desired level of accuracy, and can then be more cheaply evaluated (Beck, Friedrich, Brandani, Guillas, & Fraga, 2012; Caballero & Grossmann, 2008; Henao & Maravelias, 2010; 2011; Viana, Simpson, Balabanov, & Toropov, 2014).

Over the past years, regression analysis has been established as a powerful tool in a wide range of applications, including: customer

* Corresponding author. Tel.: +442076792563; fax: +442073882348.
E-mail addresses: lingjian.yang.10@ucl.ac.uk (L. Yang), s.liu@ucl.ac.uk (S. Liu), sophia.tsoka@kcl.ac.uk (S. Tsoka), l.papageorgiou@ucl.ac.uk (L.G. Papageorgiou).

http://dx.doi.org/10.1016/j.eswa.2015.08.034
0957-4174/

demand forecasting (Kone & Karwan, 2011; Levis & Papageorgiou, 2005), investigation of CO_2 capture process (Nuchitprasittichai & Cremaschi, 2013; Zhang & Sahinidis, 2013), optimisation of moving bed chromatography (Li, Feng, P., & Seidel-Morgenstern, 2014b), forecasting of CO_2 emission (Pan, Kung, Bretholt, & Lu, 2014), prediction of acidity constants for aromatic acids (Ghasemi, Saaidpour, & Brown, 2007), prediction of induction of apoptosis by different chemical components (Afantitis et al., 2006) and estimation of thermodynamic property of ionic liquids (Chen, Wu, & He, 2014; Wu, Chen, & He, 2014).

A large number of regression analysis methodologies exist in the literature, including: linear regression, support vector regression (SVR), kriging, radial basis function (RBF) (Sarimveis, Alexandridis, Mazarakis, & Bafas, 2004), multivariate adaptive regression splines (MARS), multilayer perceptron (MLP), random forest, K-nearest neighbour (KNN) and piecewise regressions. We briefly summarise those methodologies before presenting our proposed method.

Linear regression

Linear regression is one of the most classic types of regression analysis, which predicts the output variables as linear combinations of the input variables. The regression coefficients of the input variables are usually estimated using least squared error or least absolute error approaches. The problems can be formulated as either quadratic programming or linear programming problems, which can be solved efficiently. In some cases when the estimated linear relationship fails to adequately describe the data, a variant of linear regression analysis, called polynomial regression, can be adopted to accommodate non-linearity (Khuri & Mukhopadhyay, 2010). In polynomial regression, higher degree polynomials of the original independent input variables are added as new input variables, before estimating the coefficients of the aggregated regression function. Polynomial functions of second-degree have been most frequently used in literature due to its robust performance and computational efficiency (Khayet, Cojocaru, & Zakrzewska-Trznadel, 2008; Minjares-Fuentes et al., 2014).

Another popular variant of linear regression is called least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1994). In LASSO, summation of absolute values of regression coefficients is added as a penalty term into the objective function. The nature of LASSO encourages some coefficients to equal to *0*, thus performing implicit feature selection (Tibshirani, 2011).

Automated learning of algebraic models for optimisation (ALAMO) (Cozad, Sahinidis, & Miller, 2014; Zhang & Sahinidis, 2013) is a mathematical programming-based regression method that proposes low-complexity functions to predict output variables. Given the independent input features, ALAMO starts with defining a large set of potential basis functions, such as polynomial, multinomial, exponential and logarithmic forms of the original input variables. Subsequently an mixed integer linear programming model (MILP) is solved to select the best subset of T basis functions that optimally fit the data. The value of T is initially set equal to *1* and then iteratively increased until the Akaike information criterion, which estimates the generalisation of the constructed model, starts to decrease (Miller et al., 2014). The integer programming model is capable of capturing the synthetic effect of different basis functions, which is considered more efficient than traditional step-wise feature selection.

SVR

Support vector machine is a very established statistical learning algorithm, which fits a hyper plane to the data (Smola & Schlkopf, 2004). SVR minimises two terms in the objective function, one of which is ϵ-insensitive loss function, i.e. only sample training error greater than an user-specific threshold, ϵ, is considered in the loss function. The other term is model complexity, which is expressed as sum of squared regression coefficients. Controlling model complexity usually ensures the model generalisation, i.e. high prediction accuracy in testing samples. Another user-specified trade-off parameter balances the significance of the two terms (Bermolen & Rossi, 2009; Chang & Lin, 2011). One of the most important features that contribute to the competitiveness of SVR is the kernel trick. Kernel trick maps the dataset from the original space to higher-dimensional inner product space, at where a linear regression is equivalent to an non-linear regression function in the original space (Li, Gong, & Liddell, 2000). A number of kernel functions can be employed, e.g. polynomial function, radial basis function and fourier series (Levis & Papageorgiou, 2005). Formulated as a convex quadratic programming problem, SVR can be solved to global optimality.

Despite the simplicity and optimality of SVR, the problem of tuning two parameters, i.e. training error tolerance ϵ and trade-off parameter balancing model complexity and accuracy, and selection of suitable kernels still considerably affect its performance accuracy (Cherkassky & Ma, 2004; Lu, Lee, & Chiu, 2009).

Kriging

Kriging is a spatial interpolation-based regression analysis methodology (Kleijnen & Beers, 2004). Given a query sample, kriging estimates its output as a weighted sum of the outputs of the known nearby samples. The weights of samples are computed solely from the data by considering sample closeness and redundancy, instead of being given by an arbitrary decreasing function of distance (Kleijnen, 2009). The interpolation nature of kriging means that the derived interpolant passes through the given training data points, i.e. the error between predicted output and real output is zero for all training samples. Different variants of kriging have been developed in literature, including the most popular ordinary kriging (Lloyd & Atkinson, 2002; Zhu & Lin, 2010) and universal kriging (Brus & Heuvelink, 2007; Sampson et al., 2013).

MARS

MARS (Friedman, 1991) is another type of regression analysis that accommodates non-linearity and interaction between independent input variables in its functional relationship. Non-linearity is introduced into MARS in the form of the so-called hinge functions, which are expressions with max operators and look like $max(0, X - const)$. If independent variable X is greater than a constant number *const*, the hinge function is equal to *X-const*, otherwise the hinge function equals to *0*. The hinge functions create knots in the prediction surface of MARS. The functional form of MARS can be a weighted sum of constant, hinge functions and products of multiple hinge functions, which makes it suitable to model a wide range of non-linearity (Andrs, Lorca, de Cos Juez, & Snchez-Lasheras, 2011).

The building of MARS usually consists of two steps, a forward addition and a backward deletion step. In the forward addition step, MARS starts from one single intercept term and iteratively adds pairs of hinge functions (i.e. $max(0, X - const)$ and $max(0, const - X)$) that leads to largest reduction in training error. Afterwards, a backward deletion step, which removes one by one those hinge functions contributing insignificantly to the model accuracy, is employed to improve generalisation of the final model (Balshi et al., 2009; Leathwick, Elith, & Hastie, 2006). The presence of hinge functions also make MARS a piece-wise regression method.

MLP

Multilayer perceptron is a feedforward artificial neural network, whose structure is inspired by the organisations of biological neural networks (Hill, Marquez, O'Connor, & Remus, 1994). A MLP typically consists of an input layer of input variables, an output layer of response variables, sandwiching multiple intermediate layers of neurons. The network is fully interconnected in the sense that neurons in each layer are connected to all the neurons in the two neighbour layers (Comrie, 1997; Gevrey, Dimopoulos, & Lek, 2003). Each neuron in the intermediate layers takes a weighted linear combination of outputs from all neurons in the previous layer as input, applies an non-linear transformation function before supplying the output to all neurons of the next layer. The use of non-linear transformation functions, including sigmoid, hyperbolic tangent and logarithmic

functions, makes MLP suitable for modelling highly non-linear relationship (Gevrey et al., 2003; Rafiq, Bugmann, & Easterbrook, 2001).

Identifying the optimal configuration of a MLP, i.e. the number of intermediate layers, the number of neurons for each intermediate layer, the type of activation function for each neuron and the weights of connection between consecutive layers of neurons, is known to be time-consuming and traps in local optimal solutions (Paliwal & Kumar, 2009). The large degree of freedom in training a MLP is often blamed for data over-fitting. Dua (2010) has presented a two-objective mathematical formulation trying to find the best configuration of a MLP by balancing the training accuracy and network complexity. More often, architecture of a MLP is fixed by the user and back-propagation is used to tune only the weights of connection between neighbour layers of neurons (Gudise & Venayagamoorthy, 2003; Zhang, Zhang, Lok, & Lyu, 2007).

Random forest

Before introducing random forest we first describe *regression tree*, which is a decision tree-based prediction model. Starting from the entire set of samples, a regression tree selects one independent input variable among all and performs binary split into two child sets, under the condition that the two child nodes give increased purity of the data compared with its single parent node. Purity is often defined as the deviation of predicting with the mean value of the output variable. The process of binary split is recursively applied for each child node until a terminating criterion is satisfied. The nodes that are not further partitioned are called leaves. After growing a large tree, a pruning process is employed to remove the leaves contributing insignificantly to the purity improvement (Breiman, Friedman, Olshen, & Stone, 1984; Loh, 2011). In order to improve model fit, a linear regression model can be fitted for each leaf (Quinlan, 1992).

Random forest is an ensemble learning method of regression trees. In general, random forest (Biau, 2012; Breiman, 2001) builds a forest of multiple regression tree models and aggregate the decisions from all the trees to produce a final prediction. Given a dataset, multiple bootstrap sample sets are first created by random sampling with replacement. Each of the bootstrap sample set is then learned by a revised regression tree algorithm, which differs from the classic regression tree by randomly selecting a candidate subset of features for each binary split of node (Genuer, Poggi, & Tuleau-Malot, 2010). The accuracy of each regression tree can be estimated on the training samples absent from the bootstrap set, and the final prediction can be either simple average of predictions from all trees or weighted average considering the estimated accuracy. It is demonstrated that random forest achieves much robust prediction performance compared with single regression tree method (Breiman, 2001; Fanelli, Gall, & Van Gool, 2011).

KNN

KNN belongs to the category of lazy learning algorithms, due to the fact that prediction is based on the instances without an explicit training phase of constructing models, thus making it one of the simplest regression methods in literature (Korhonen & Kangas, 1997). Given an enquiry sample, KNN first identifies K closest instances in the training sample set, the exact value of K is given *a priori*. The closeness of samples can be measured by different distance metrics, for example Euclidean and Manhattan distances (Eronen & Klapuri, 2010; Scheuber, 2010). Prediction is then taken as weighted mean of the outputs of the K nearest neighbours, with weight often being defined as the inverse of distance (Papadopoulos, Vovk, & Gammerman, 2011). Despite its simplicity, KNN usually provides competitive prediction performance against much more sophisticated algorithms.

Previous work on piecewise regression

Piecewise functions have been frequently studied in literature as well. In (Toms & Lesperance, 2003), univariate piece-wise linear functions have been used to fit ecological data and identify break-points that represent critical threshold values of a phenomenon. In (Strikholm, 2006), a method based on statistical testing is proposed to estimate the number of break-points for an univariate piece-wise linear function. Malash and El-Khaiary (2010) also apply piece-wise linear regression techniques on univariate experimental adsorption data. Piece-wise function is determined by solving a non-linear programming model. SegReg (www.waterlog.info/segreg.htm) is free software that permits estimating of piece-wise regression functions with up to two independent variables. For one independent variable, SegReg splits from a series of candidate break-points and for each one fits a linear regression for either side of the break-point. The break-point corresponding to the largest statistical confidence is taken as the final solution. In the case of two independent variables, SegReg first determines the two-region piece-wise regression function between the dependent variable and the most significant input variable, before computing the relation between its residual/deviation and the second input variable.

Both Magnani and Boyd (2009) and Toriello and Vielma (2012) publish work on data fitting with a special family of piece-wise regression functions, called max-affine functions. The form of max-affine functions is defined as the maximum of a series of linear functions, i.e. a sample is projected to all linear functions, and the maximum projected value among all is taken as final predicted value from the piece-wise functions. The use of max-affine functions limits the fitted surface to be convex. In (Magnani & Boyd, 2009) a heuristic method is used to ease the difficulty of direct solving the highly non-linear max-affine functions, while in (Toriello & Vielma, 2012), big-M constraint is used to reformulate the problem into an non-convex mixed integer non-linear programming model. However, computational complexity is limiting their applications to examples of small scale.

More recently, Greene, Rolfson, Garellick, Gordon, and Nemes (2015) applies piece-wise regression analysis to predict patient's post-treatment life quality with the pre-treatment life quality measure, which identifies the segments where therapy benefits vary significantly. The analysis is performed using Segmented (Muggeo, 2008), a package written in R (R Development Core Team, 2008). Segmented formulates the problem using a non-linear model and requires a user to specify the segmented input variables, the number of break-points and also the initial guess of each break-point. Starting from the those user supplied initial positions of break-points, Segmented iteratively moves around the neighbour of the initial guess points to search break-points of better quality using local linearisation. However, it is difficult if not impossible to reasonably guess good starting points for real world multivariate problems of large number of samples and input variables, where visual examination cannot be performed. This makes it hard to identify quality solutions. Furthermore, Segmented only allows the input variables being partitioned to have different regression coefficients across different segments, while the other input variables keep the same coefficients within the entire ranges, significantly restricting its flexibility.

In both (Xue, Liu, Zhang, & Hu, 2013) and (Li, Zhang, Yan, & Xue, 2014a), piece-wise regression function were employed to detect vegetation changes. Piece-wise linear regression was tackled using fuzzy logic and identifies the changes in patterns of vegetation greenness. Cavanaugh et al. (2014) employ piece-wise regression and find out that the changes in mangrove area over the last 20 years is a piecewise functions of latitude, with regions above and below a specific threshold latitude value following two different patterns of mangrove grows. Moreover, Matthews, Steinbauer, Tzirkalli, Triantis, and Whittaker (2014) uses 2-segment piece-wise functions to describe the relationship between species richness and fragment area of islands, with the critical breakpoint being determined by simply sampling a number of candidate values and selecting the one giving best model fit. Unfortunately, the above methods are all limited to model rather simple relationship between one output variable and one input variable, seriously limiting their usage in more complex problems.

It is clear that the previous literature work of piece-wise regression methods are non-linear and computationally restricted to problems of very small scales. Yet, they cannot be solved to identify globally optimal solutions. In this work, we propose a novel linear model for piece-wise regression analysis. A single input variable is partitioned to separate samples into multiple mutually exclusive segments, while each segment is fitted with a unique multivariate linear regression function. Assuming that both the partition variable and the number of break-points are known, we propose an optimisation model that optimally estimates the position of all break-points and the linear regression coefficients for each segment *simultaneously* so that the total absolute deviation is minimised. Thanks to the usage of binary variables, our proposed mathematical model is linear and can be efficiently solved to global optimality for problems up to thousands samples (see Section 3). Furthermore, a solution procedure is used to identify the key partition variable and the final number of break-points. Several real world multivariate benchmark datasets have been used to demonstrate the applicability and efficiency of the proposed method.

The proposed piece-wise regression method can help construct expert systems in various application domains. Expert systems are computer programs designed to make decisions analogous to human experts. As an expert system is typically made up of an inference engine and a knowledge base, the quality and quantity of information in knowledge base directly affects the usefulness of the constructed expert system. Our proposed piece-wise regression method can be helpful in more efficiently building expert systems via automatic and efficient acquisition of knowledge. More specifically, the proposed piece-wise regression method can extract latent knowledge from the large collection of domain expert curated databases. Those discovered knowledge are represented in the form of identified relationship between input and output variables of interest, which can be combined with expert knowledge to form the final expert system (Alonso, Martnez, Prez, & Valente, 2012). For example, the proposed piece-wise regression method in this work can be used for building prognostic expert systems in medical applications. When presented historical data of patients' clinical variables and survival length, piece-wise regression can induce domain knowledge by approximating the complex relationship between clinical variables and survival length. Those induced knowledge can then be used to perform prognosis for the current patients, imitating the end-behaviour of human experts, i.e. medical doctors.

Overall, the key contributions of our work are illustrated below:

- We propose a novel mixed integer optimisation model for multivariate regression analysis modelling piece-wise linear functions, which partitions a single variable into multiple mutually exclusive regions and fits each one with a distinct multivariate linear function. Given *a priori* a single input variable as partition feature and the number of segments, the optimisation model can be solved to simultaneously determine the positions of multiple break points and regression coefficients for each segment.
- Given that neither which feature should be segmented nor the number of segments are typically known, a heuristic solution procedure is also introduced that automatically identifies the key partition variable and the final number of segments.
- A number of real world benchmark problems have been employed to demonstrate the applicability and efficiency of the proposed method. As sharp difference to the existing piece-wise regression methods in literature which can only be applied to problems of very small size, our proposed optimisation model can be solved to global optimality for datasets containing up to five thousand samples. Comparison to some popular regression methods based on other methodologies clearly indicates that our proposed method based on piece-wise function achieves the highest prediction accuracy, and does it consistently. Besides high prediction performance, our proposed regression method has the advantage of being easily understandable and interpretable, as the learned model can be conveniently represented as a small set of rules.
- As a generic data mining method, our proposed regression method can help with constructions of expert and intelligent systems via automatic extraction of knowledge from database. We will discuss its potential usage in various application domains.

The rest of the paper is structured as follows: in Section 2, we present the mathematical programming model and a heuristic solution procedure. In Section 3, comparative results of our proposed method and some state-of-the-art regression algorithms on benchmark examples are presented and discussed. The last section concludes with our key findings.

2. Method

A novel piecewise linear regression method is proposed in this work. The core idea of the proposed method is to identify a single input feature, and separate the samples into complementary regions on this feature. One different linear regression function is fitted locally for each region. The sample partition and calculation of local regression coefficients are performed simultaneously within the proposed optimisation to achieve least absolute error.

2.1. A novel regression method

In this section, we first describe a novel mathematical programming model that optimises the location of break-points and regression coefficients for each region so as to achieve minimal training error. Subsequently, a solution procedure is proposed to identify the best partition feature and the number of regions.

The indices, parameters and variables associated with the proposed model are listed below:

Indices

s	*sample*, $s = 1, 2, \ldots, S$
m	*feature/independent input variable*, $m = 1, 2, \ldots, M$
r	*region*, $r = 1, 2, \ldots, R$
m^*	*the feature where sample partition takes place*

Parameters

A_{sm}	*numeric value of sample s on feature m*
Y_s	*output value of sample s*
U', U''	*arbitrarily large positive numbers*

Continuous variables

W_m^r	*regression coefficient for feature m in region r*
B^r	*intercept of regression function in region r*
$Pred_s^r$	*predicted output for sample s in region r*
$X_{m^*}^r$	*break-point r on partition feature* m^*
D_s	*training error between predicted output and real output for sample s*

Binary variables

F_s^r	1 *if sample s falls into region r*; 0 *otherwise*

Assume first that both the partition feature m^* and the number of regions R are given, the *R-1* break points are arranged in an ordered way:

$$X_m^{r-1} \le X_m^r \quad \forall m = m^*, \quad r = 2, 3, \ldots, R \tag{1}$$

Binary variables F_s^r are introduced to model if sample s belongs to region r or not. Modelling of which sample belongs to which region is achieved with the following constraints:

$$X_m^{r-1} - U'(1 - F_s^r) \le A_{sm} \quad \forall s, r = 2, 3, \ldots, R, m = m^* \tag{2}$$

$$A_{sm} \le X_m^r + U'(1 - F_s^r) \quad \forall s, r = 1, 2, \ldots, R-1, m = m^* \tag{3}$$

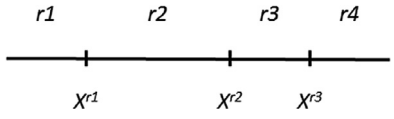

Fig. 1. Break-points and regions.

When sample s belongs to region r (i.e. $F_s^r = 1$), A_{sm^*} falls into the region bounded by the two consecutive break-points $X_{m^*}^{r-1}$ and $X_{m^*}^r$ on feature m^*; otherwise the two sets of constraints become redundant. A visualisation of break-points and regions is provided in Fig. 1:

The following constraints restrict that each sample belongs to one and only one region:

$$\sum_r F_s^r = 1 \quad \forall s \tag{4}$$

For sample s, its predicted output value for region r, $Pred_s^r$, is as below:

$$Pred_s^r = \sum_m A_{sm} W_m^r + B^r \quad \forall s, r \tag{5}$$

For any sample s, its training error is equal to the absolute deviation between the real output and the predicted output for the region r where it belongs to (i.e. $F_s^r = 1$):

$$D_s \geq Y_s - Pred_s^r - U''(1 - F_s^r) \quad \forall s, r \tag{6}$$

$$D_s \geq Pred_s^r - Y_s - U''(1 - F_s^r) \quad \forall s, r \tag{7}$$

The objective function is to minimise the sum of absolute training error:

$$min \sum_s D_s \tag{8}$$

The final model, named as Optimal Piece-wise Linear Regression Analysis (*OPLRA*) in this work, consists of a linear objective function and several linear constraints, and the presence of both binary and continuous variables define an MILP problem, which can be solved to global optimality by standard solution algorithms, for example branch and bound. A heuristic solution procedure is also employed in this work to identify the partition feature and the number of regions, as described in Fig. 2 below.

The heuristic procedure starts with solving a linear regression on the entire set of data with least absolute deviation. Subsequently, each input feature in turn serves as partition feature m^* once and the *OPLRA* model is solved while allowing two regions (i.e. $R = 2$). The feature corresponding to the minimum training error is kept and if its error represents a percentage reduction of more than β from the global linear regression without data partition, the procedure continues; otherwise it is decided that two-region piecewise linear regression does not provide a desirable improvement upon the classic linear regression, and the initially derived linear regression function without sample partition is obtained for prediction. The user-specified parameter β, taking value between *0* and *1*, quantifies the percentage reduction in training error that justifies adding one more region. If two-region piecewise regression is accepted, the

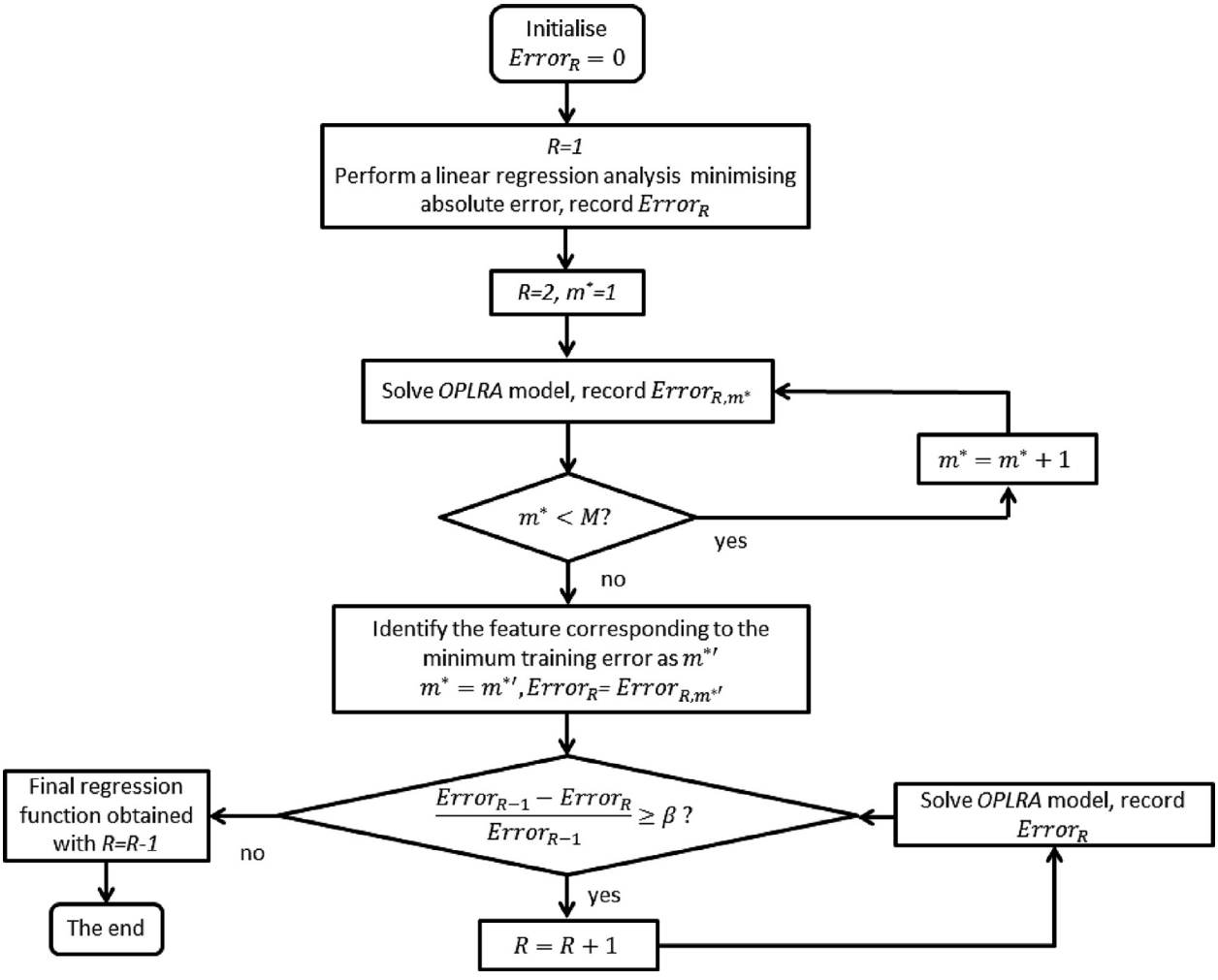

Fig. 2. Heuristic procedure to identify the partition feature and the number of regions.

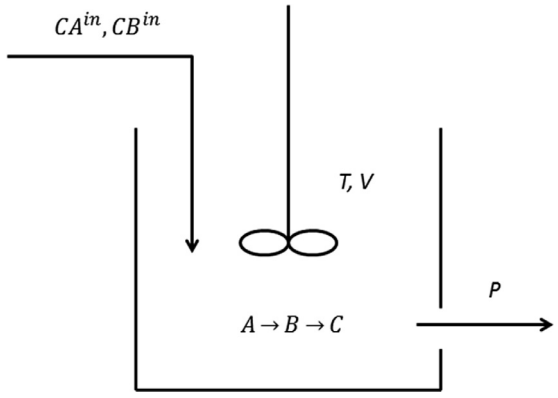

Fig. 3. Illustrative example of a continuous stirred tank reactor.

corresponding partition feature is retained for further analysis while the number of regions is iteratively increased, until the β training reduction criterion is not satisfied between iterations.

β is the only user-specific parameter in our proposed regression method, which requires fine tuning. A small value may cause over-fitting, i.e. too many regions are allowed and each region contains only a small number of samples, which then results in unreliable construction of local linear regression functions; while a value excessively large will lead to premature growing of regions, which then under-fit the data. In Section 3, we will test a series of values on a number of benchmark datasets and select the optimal value corresponding to the most robust prediction performance.

The constructed piecewise linear regression functions are then used to predict the output value of new samples. A testing sample is firstly assigned to one of the regions, and the regression coefficients for that region are used to estimate its output value.

2.2. An illustrative example

In order to better illustrate the training of the proposed regression method, a simulation model is taken from literature. In brief, the illustrative example (Palmer & Realff, 2002) describes the operation of a continuous stirred tank reactor, where a chain reaction of $A \rightarrow B \rightarrow C$ takes place. An inlet stream containing both reactant A and B enters the reactor and the desirable output is component B. There are 4 independent input variables to the simulation model, including temperature of the reactor (T), volume of the reactor (V), concentration of A and B in the inlet stream (CA^{in} and CB^{in}). The output to be predicted is the production rate of B (P). The process and associated variables are described in Fig. 3.

With latin hypercube sampling technique (Helton & Davis, 2003) employed to specify a set of data points, we run the simulation model and collect 300 samples. The goal of the regression analysis is to approximate the functional relationship between output variable P and input variables including T, V, CA^{in} and CB^{in} using piece-wise linear functions. The step-wise description of the training procedure is presented in Table 1 below.

Initially, a linear regression function is fitted to the entire dataset without feature segmentation, which gives an absolute deviation of 1677.78. The second iteration of the method solves 4 independent *OPLRA* models allowing 2 regions each, respectively specifying T, V, CA^{in} and CB^{in} as partition feature. The two-region piece-wise linear functions constructed while partitioning on T appears to yield lower training errors (i.e. 1030.63) than the other 3, and therefore is taken as the solution for iteration 2. This represents a significant improvement (i.e. 38.57%) from the initial global linear regression function. From iteration 3, the partition feature is fixed as T while one more region is allocated for each increased iteration. Iteration 3 and 4 respectively lowers the training error to 876.66 and 807.12. The iterative procedure terminates when the β criterion is not satisfied, e.g. if $\beta = 20\%$, then the iterative procedure terminates at the third iteration and the final regression function has 2 regions; if $\beta = 10\%$, then the final regression function has 3 regions.

Overall, the key features of our proposed piecewise linear regression method are summarised here: (1) our method identifies one key partition feature and separate samples into multiple complementary regions on it, (2) each region has the flexibility of being fitted by its own linear regression function, with all input features allowed to have different regression coefficients across different regions, (3) there is only one tuning parameter β, (4) compared with algorithms like kernel-based SVR and MLP, the constructed regression function is easy to understand, as it exhibits linear relationships for different regions.

It is noted here that the obtained relationship between input and output variables, presented as rules in Table 1, can be used to build an expert system for the above operation. Given the chain reaction

Table 1
Piecewise regression functions built at each step of training procedure.

Iteration	Number of regions	Partition feature	Training error	Training error improvement	Functional relationship
1	1	NONE	1677.78		$P = 1.0240T + 0.0054CA^{in} + 0.0125CB^{in} + 0.4340V - 333.54$
2	2	T	1030.63	38.57%	$P = \begin{cases} 0.7413T + 0.0040CA^{in} + 0.0102CB^{in} + 0.3406V - 238.74, & T \le 213.21 \\ 1.7156T + 0.0111CA^{in} + 0.0315CB^{in} + 0.7574V - 592.63, & T > 213.21 \end{cases}$
	2	V	1143.49		$P = \begin{cases} 0.5952T + 0.0033CA^{in} + 0.0056CB^{in} + 0.4533V - 194.26, & V \le 42.38 \\ 1.4781T + 0.0083CA^{in} + 0.0195CB^{in} + 0.4773V - 48.70, & V > 42.38 \end{cases}$
	2	CA^{in}	1485.65		$P = \begin{cases} 0.8930T + 0.0057CA^{in} + 0.0152CB^{in} + 0.4161V - 293.45, & CA^{in} \le 3528.43 \\ 1.4857T + 0.0073CA^{in} + 0.0070CB^{in} + 0.5929V - 489.45, & CA^{in} > 3528.43 \end{cases}$
	2	CB^{in}	1627.73		$P = \begin{cases} 1.0242T + 0.0056CA^{in} + 0.0118CB^{in} + 0.4241V - 333.49, & CB^{in} \le 458.21 \\ 1.1105T + 0.0050CA^{in} - 0.1405CB^{in} + 0.5813V - 291.00, & CB^{in} > 458.21 \end{cases}$
3	3	T	876.66	14.94%	$P = \begin{cases} 0.5815T + 0.0030CA^{in} + 0.0097CB^{in} + 0.2654V - 184.45, & T \le 303.25 \\ 1.1353T + 0.0062CA^{in} + 0.0176CB^{in} + 0.4579V - 373.68, & 303.25 < T \le 316.62 \\ 1.8764T + 0.0119CA^{in} + 0.0394CB^{in} + 0.8617V - 654.41, & T > 316.62 \end{cases}$
4	4	T	807.12	7.93%	$P = \begin{cases} 0.5815T + 0.0030CA^{in} + 0.0097CB^{in} + 0.2654V - 184.45, & T \le 303.25 \\ 1.2648T + 0.0054CA^{in} + 0.0148CB^{in} + 0.4510V - 409.61, & 303.32 < T \le 312.21 \\ 1.4872T + 0.0084CA^{in} + 0.0202CB^{in} + 0.6667V - 503.10, & 312.21 < T \le 320.77 \\ 1.9930T + 0.0128CA^{in} + 0.0360CB^{in} + 0.8871V - 695.65, & T > 320.77 \end{cases}$
...					

of $A \rightarrow B \rightarrow C$ in stirred tank reactor (Palmer & Realff, 2002), domain experts perform experiments to create a database of samples for different levels of temperature, reactor volume and concentrations of reactants. Our proposed piece-wise regression method is then applied to automatically extract the rules that predict production rate from temperature, reactor volume and reactant concentrations. The rules will be difficult to be provided directly from even chemical engineering experts due to the complex nature of the reaction. Since, the above extracted rules can calculate a production rate value for any random value of temperature, tank volume and reactant concentrations, regardless of if they obey physical laws (must be positive) or are valid for the reaction of interest, expert knowledge should be incorporated to further refine the rules. For example, expert knowledge can be used to constraint the applicable temperature range outside which liquid phase will vaporise to gas phase or freeze to solid phase, making it impossible for the reaction to proceed as normal. The final expert system will allow users to query the likely outcome, as production rate or no reaction, of any combination of values of temperature, reactor volume and reactant concentrations.

In the next section, a number of real world regression problems are employed to benchmark the predictive performance of our proposed model.

3. Results and discussion

A total number of 7 real world datasets have been downloaded from UCI machine learning repository (http://archive.ics.uci.edu/ml/) (Bache & Lichman, 2013) to test the prediction performance of our proposed method. The first regression problem Yacht Hydrodynamics predicts the hydrodynamic performance of sailing yachts from 7 features describing the hull dimensions and velocity of the boat for 308 samples. Energy Efficiency (Tsanas & Xifara, 2012) collects data corresponding to 768 building shapes, described by 8 features including wall area, root area and so on. The aims are to establish the relationship between either heating load or cooling load requirements and the 8 parameters of the building. The third example, Concrete Strength (Yeh, 1998), looks into the relationship between compressive strength of concrete and 8 input variables, including water concentration and age, with 1030 samples of different concretes. Airfoil dataset concerns how the different airfoil blade designs, wind speed and angles of attack affect the sound pressure level. The last 2 case studies, Red Wine Quality and White Wine Quality (Cortez et al., 2009), aims to predict experts' preference of red and white wine taste with 11 physicochemical features of the wines. Almost 1600 red wine and 4900 white wine samples have been obtained for analysis.

For each of the 7 benchmark datasets, 5-fold cross validation, is performed to estimate the predictive accuracy of the proposed method. Given a dataset, 5-fold cross validation randomly splits the samples into 5 subsets of equal size. Each subset is in turn held out once while the other 4 subsets of samples are used in the training process to derive the regression function. The holdout set is then used to validate the predictive accuracy of the constructed regression function. We conduct 10 rounds of 5-fold cross validation by performing different random sample splits, and the mean absolute prediction errors (MAE) are averaged over 50 testing sets as the final error.

For comparison purposes, a number of state-of-the-art regression methods have been implemented, including linear regression, MLP, kriging, SVR, KNN, random forest, MARS, PaceRegression and ALAMO. Linear regression, MLP, kriging, SVR, KNN and PaceRegression are implemented in WEKA machine learning software (Hall et al., 2009). For KNN, the number of nearest neighbours is selected as 5, while for other methods their default settings have been retained. Random forest is implemented using Orange (Demšar et al., 2013). We use a MATLAB toolbox called ARESlab (Jekabsons, 2011) for MARS. ALAMO is reproduced using the General Algebraic Modelling System (GAMS) (GAMS Development Corporation, 2013), and basis function forms including polynomial of degrees up to 3, pair-wise multinomial terms of equal exponents up to 3, exponential and logarithmic forms are provided for each dataset. Our proposed method is also implemented in GAMS. Both ALAMO and our proposed model are solved using CPLEX MILP solver, with optimality gap set as 0. Computational resource limit is set as 200 seconds for each solving of *OPLRA* model in our proposed method.

3.1. Sensitivity analysis for β

In this subsection, a sensitivity analysis is performed for the parameter β, which serves as a terminating criterion of the iterative training procedure for our proposed method. Taking value between *0* and *1*, β defines the minimum percentage training error reduction that must be achieved to justify the allocation of an extra region. A range of values have been tested, including: *0.2, 0.15, 0.10, 0.05, 0.03* and *0.01*. The results of the sensitivity analysis are provided in Fig. 4 below.

Fig. 4 describes how mean absolute error changes with β. The numbers attached to the points in each plot are the average numbers of final regions, which always go up as β decreases. For Yacht Hydrodynamics example, setting $\beta = 0.20$ results in just more than 4 final regions. Decrease the β value to *0.15* increases slightly the prediction error with marginally higher number of regions. Further decrease β to *0.10* leads to lowest mean prediction error of 0.648 with an average of 5 regions, before excessively low values of β over-fits the unseen testing samples by yielding much increased prediction error. For Energy Efficiency Heating case study, when $\beta = 0.10, 0.15$ and *0.20* our proposed regression method constructs piece-wise regression functions of an average of 3 regions, yielding MAE of 0.907. Smaller values of β leads to about 5 regions, which are shown to predict the testing samples with higher accuracy (MAE around 0.810). In terms of Energy Efficiency Cooling and Concrete Strength examples, similar phenomenon can be observed that when β takes overly high values (i.e. *0.20, 0.15*), the proposed method terminates prematurely with only 2 regions and relatively high MAE. More regions are allowed by lowering β, which gives higher prediction accuracies. On Airfoil case study, the proposed method outputs global multiple linear regression functions without data partitions when $\beta = 0.20$. As β decreases, more regions are permitted, which predict unseen samples with better accuracy. With regards to Red Wine Quality dataset, the optimal prediction occurs when $\beta = 0.03$. On the last example of White Wine Quality, 2-region piece-wise regression functions achieved with $\beta = 0.01, 0.03, 0.05$ outperforms global multiple linear regressions for higher values of β.

It can be seen from Fig. 4 that the range of values between *0.01* and *0.05* generally lead to smaller prediction error than higher values of β. For all datasets except Yacht Hydrodynamics, prediction errors of $\beta = 0.01, 0.03$ and *0.05* are evidently smaller than that of $\beta = 0.10, 0.15$ and *0.20*. Within the range between *0.01* and *0.05*, there is no clear optimal value for β as different values have different effects on the accuracy. We instead seek to identify the most robust value for β, which gives consistently desirable prediction accuracy across a wide range of problems. For each dataset, we normalise the MAE of each β according to the formula: $\frac{MAE_\beta - \min_\beta MAE_\beta}{\min_\beta MAE_\beta}$. For example, in Yacht Hydrodynamics, original MAE of $\beta = 0.01$ is normalised from 0.7131 to $\frac{0.7131-0.6481}{0.6481} = 10.0\%$, where 0.6481 is the lowest MAE achieved when $\beta = 0.10$. The normalised MAE of each β represents the actual deviation of it compared to the lowest error, and is averaged over all examples to reflect its overall competitiveness.

Overall $\beta = 0.03$ provides the smallest normalised MAE of 1.7%, which is marginally lower than these of $\beta = 0.01$ and $\beta = 0.05$, respectively as 1.8 and 2.8%. Even higher values of β correspond to noticeably larger normalised MAE (5.6, 9.7 and 12.3% for $\beta = 0.10, 0.15$ and *0.20*, respectively). The consistently small normalised MAE, while β is between *0.01* and *0.05*, show that our proposed regression

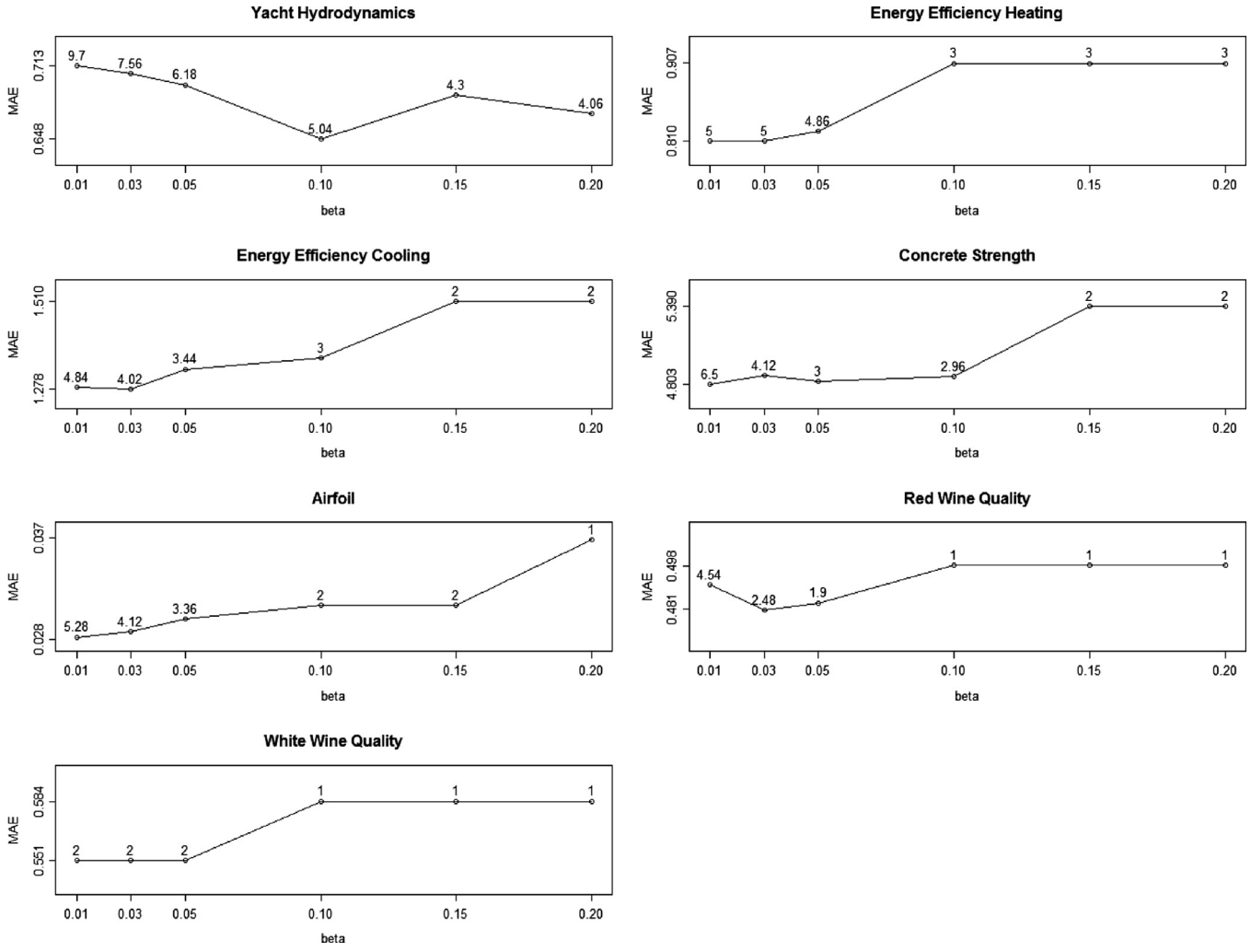

Fig. 4. Sensitivity analysis of β. The numbers above points in each plot correspond to the average numbers of final regions.

method is robust with respect to the only user tuning parameter β. Finally β is set to *0.03* when comparing with other competing methods in literature.

3.2. Prediction performance comparison

After identifying a value (i.e. *0.03*) for the only tuning parameter β in our proposed regression method, we now compare the accuracy of the proposed method against some popular regression algorithms with the same set of 7 examples. The results of the comparison are available in Table 2 below.

In Table 2 and each tested dataset, the lowest prediction error achieved among all implemented regression methods is marked with bold. On Hydrodynamics problem, the proposed method in this work provides an MAE of 0.706, which is lower than any other competing algorithm. ALAMO, MLP and MARS follow closely with MAE of 0.787, 0.809 and 1.011, respectively. Mean error rates of the rest of the methods are between 3 and 8. On Energy Efficiency Heating, MARS emerges as the most accurate algorithm with an mean absolute error of 0.796, which is closely matched by our proposed method and MLP. Mean prediction errors of the other approaches are almost all twice as large as that of the MARS. In terms of Energy Efficiency Cooling dataset, the proposed method, MARS, random forest and MLP are the top 4 performers with MAE between 1.278 and 1.924. On Concrete Strength, our proposed approach and MARS, with an MAE of 4.870 and 4.871, again emerge as the leading methods from random forest, Kriging, MLP and the others. When it comes to Airfoil example, all the competing algorithms achieve similar prediction accuracies, with KNN topping the chart with an MAE of 0.026. The proposed approach in this work is a merely 0.003 far behind, with kriging and random forest a further 0.001 behind. A merely 0.011 separates the 10 methods. Lastly, on the two Wine Quality examples, our proposed approach is respectively ranked as 1st and 3rd best method.

Overall, for 4 out of the 7 datasets, including Yacht Hydrodynamics, Energy Efficiency Cooling, Concrete Strength and Red Wine Quality, the proposed regression method achieves the lowest prediction errors. For the other 3 tested examples, the proposed method still perform competitively as being second on Energy Efficiency Heating, Airfoil and third on White Wine Quality.

As there does not exist a single regression method which can always outperform others on all datasets, a desirable regression algorithm should demonstrate consistently competitive prediction accuracy. In order to more comprehensively evaluate the relative competitiveness of all the implemented approaches, we employ the following scoring strategy: for each problem, the regression methods are ranked in descending order according to their mean prediction error. The best regression method corresponding to the lowest prediction error is awarded the maximum score of 10, the second best regression method corresponding to the second lowest prediction error is assigned a score of 9 and so on. The scores of each regression approach are averaged over the 7 datasets, which represent the overall performance of the method. The higher the score, the better the relative performance of a method. The scores of the different regression approaches used in this work are presented in Fig. 5 below.

Table 2
Comparative testing of different regression methods on benchmark datasets.

	Hydrodynamics	Energy Heating	Energy Cooling	Concrete	Airfoil	Red Wine	White Wine
linear regression	7.270	2.089	2.266	8.311	0.037	0.506	0.586
MLP	0.809	0.993	1.924	6.229	0.035	0.581	0.623
Kriging	4.324	1.788	2.044	6.224	0.030	0.496	0.576
SVR	6.445	2.036	2.191	8.212	0.037	0.500	0.585
KNN	5.299	1.937	2.148	7.068	**0.026**	0.515	0.537
Random forest	3.516	1.435	1.644	5.309	0.030	0.484	**0.519**
MARS	1.011	**0.796**	1.324	4.871	0.035	0.502	0.570
PaceRegression	7.233	2.089	2.261	8.298	0.037	0.507	0.586
ALAMO	0.787	2.722	2.765	8.044	0.032	0.594	0.639
Proposed	**0.706**	0.810	**1.278**	**4.870**	0.029	**0.481**	0.551

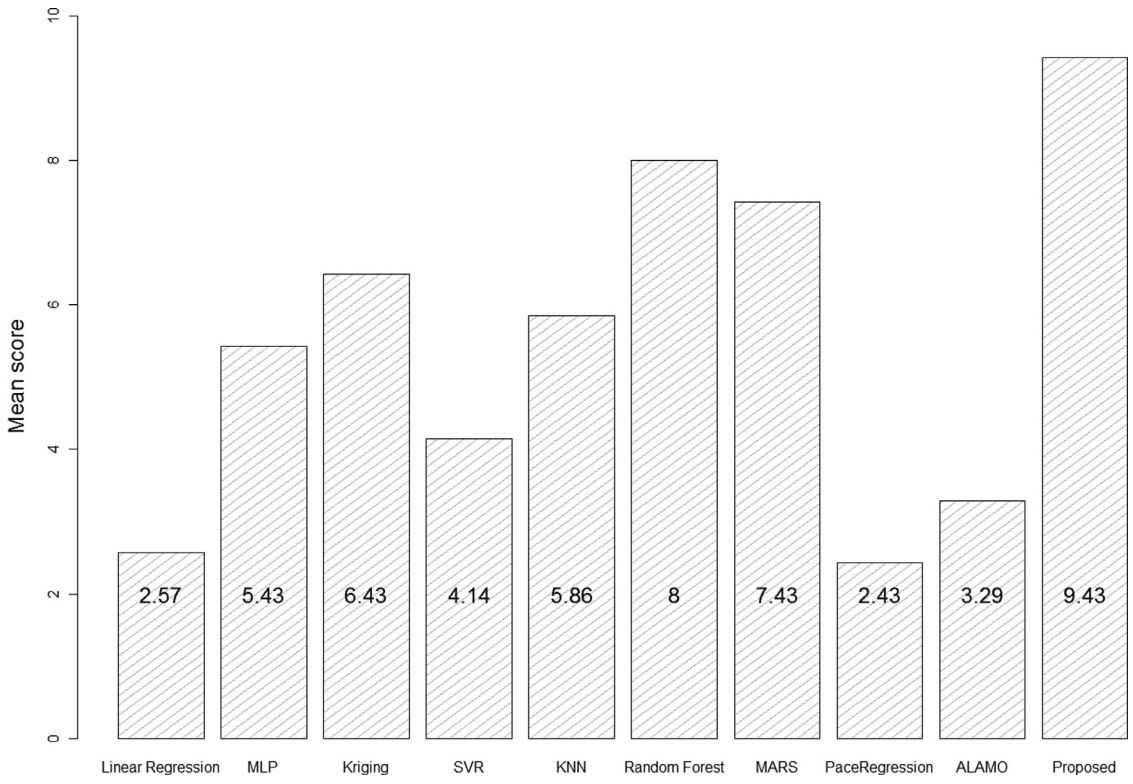

Fig. 5. Scoring of regression methods.

Table 3
Number of regions and partition feature by our proposed method.

Dataset	Number of regions	Partition feature
Hydrodynamics	5	Froude number
Energy Heating	3	Wall Area
Energy Cooling	3	Wall Area
Concrete	3	Age
Airfoil	4	Frequency
Red Wine	2	Alcohol
White Wine	2	Volatile acidity

According to Fig. 5, the proposed method is shown to be the most accurate and robust regression algorithm among all, achieving a score of 9.43 out of a possible 10. Random forest and MARS are second and third according to the ranking with scores of 8 and 7.43, followed by kriging, KNN, MLP, SVR, ALAMO, linear regression and PaceRegression in descending order. The advantages of the proposed regression method is quite obvious compared with other implemented methods.

Lastly, we take a look at, for each dataset, the number of regions and the key partition feature determined by our proposed regression method. The results are summarised in Table 3. It is clear that the proposed segmented regression method provides good interpretability as the number of regions are small (usually between 2 to 4 and at most 5). The partition features may release important insights of the underlying system as the output variables change more dramatically across different ranges alone this feature.

3.3. Strength and weakness of the proposed piece-wise regression method

No regression method will be the best for all problems. In this section, we give some general illustration of the pros and cons of the proposed *OPLRA* piece-wise linear regression method, and compare it against some other literature methods. *OPLRA* piece-wise regression is inherently deterministic, which means the same solution is always guaranteed regardless of the number of runs executed. This is an advantage of *OPLRA* against stochastic-based methods, for example MLP, where each execution would typically end up with a different locally optimal solution. On the other hand, *OPLRA* is intuitive and easy to interpret. *OPLRA* approximates the potentially highly non-linear relationship between output and input variables as piece-wise linear algebraic functions, the formalism of which is easy to understand, interpret and use for users without sophisticated background knowledge. Contrarily, the mechanisms of certain methods like SVR, MLP and Kriging lack transparency as the former two work as

black box techniques and the latter requires detailed knowledge on statistics. The small number of user-specified parameters involved in training of *OPLRA* is another remarkable advantage. β is the only tuning parameter in the proposed *OPLRA*, which produces robust predictive performance with regards to varying values of β as shown in the following Section 3. Conversely, usage of certain regression methods, including SVR, MLP and Kriging requires tuning a large number of parameters, making it a challenging task to identify their optimal values. More importantly, *OPLRA* piece-wise regression achieves more accurate and robust prediction performance against other methods. Using a large number of real word problems, *OPLRA* is shown to outperform popular state-of-the-art multivariate regression methods in terms of prediction accuracy and does so consistently across a number of real world problems.

With regards to shortcomings of our proposed piece-wise regression method, training of *OPLRA* generally consumes more computational resource than the existing methods in literature. Solving *OPLRA* combinatorial optimisation model is indeed a more computationally intensive task than heuristic-based methods, for example regression tree, MARS and quadratic programming-based SVR. Therefore, we note here that this method is not designed for online applications where computation time is valued more than the prediction accuracy/model interpretability. Another limitation of *OPLRA* is that it permits segmentation of only one input variable, which may not be adequate for the datasets where trend of the output variable changes dramatically in more than one input variables.

4. Concluding remarks

This work addresses the problem of multivariate regression analysis, where one seeks to estimate the complex relationship between dependent output variables and independent input variables from training samples. The identified relationships can then be used to make predictions for unseen observations. We have proposed a novel piece-wise regression method, which approaches the problem by segmenting one input variable into multiple mutually exclusive regions and simultaneously fitting each one with a distinct multivariate linear function. An optimisation model has been proposed to optimise the locations of break points and regression coefficients for each region, while a heuristic procedure has also been introduced to find the key partition feature and the number of break-points by repeatedly solving the optimisation models until a satisfactory solution is identified.

To demonstrate the applicability and efficiency of the proposed piece-wise regression method, 7 real world problems have been employed, covering a wide range of application domains. To benchmark the predictive capability of the proposed method, we have also implemented various popular regression methods in literature for comparison, including support vector regression, artificial neural network, MARS and K nearest neighbour. Computational experiments clearly indicate that our proposed piece-wise regression method achieves consistently high predictive accuracy as leading to the lowest prediction errors for 4 out of 7 datasets, second lowest errors for 2 datasets and third lowest error for the other example. The results confirm our proposed method as a reliable alternative to traditional regression analysis methods. Another remarkable advantage of our proposed method is that the learned model can be conveniently expressed as a set of if-then rules that are compact and easily understandable. From Table 3, it is clear that the number of if-then rules identified by our method as the hidden patterns in the large scale databases (up to thousands expert curated samples) are extremely small (usually 2 to 3 and at most 5). The model interpretability of the proposed piecewise regression is a desirable advantage over black modelling techniques, for example support vector regression and neural network.

With regards to research contribution in expert and intelligent systems, the generic machine learning method proposed in this work can be used to construct a large number of automatic decision making or support systems for various domain applications. As the quality and coverage of information contained in knowledge base critically affects the efficiency of any expert and intelligent system, our proposed machine learning method can serve to automatically and more efficiently acquire knowledge from database by approximating the relationship between output and input variables as rules. Subsequently, the discovered knowledge can be used to generate forecasts to users' enquiry.

To further improve the efficiency of the proposed piece-wise regression method in this work, the following limitations can be considered for refinement. As the piece-wise regression method proposed in this work can only partition a single input variable, one potential improvement is to generalise the method so that to permit segmentation of multiple variables so as to better capture the non-linearity in datasets. Secondly, as our proposed method in this work can only handle continuous input variables, we plan to improve its applicability by generalising it to deal with categorical input variables having many distinct levels. In addition, the relationship between output and input variables are approximated as linear for each segment in the current method, which may not adequately model the underlying patterns. To overcome this, more complex non-linear basis functions, for example polynomial, exponential and logarithmic forms, can be added to allow more flexibility. Another limitation of our method is the relatively high computational cost, which may restrict its usage in certain online applications, where learning speed of the method is considered more important than actual prediction accuracy. To tackle this problem, we can explore more efficient heuristic solution procedures that, by estimating the possible break-point positions and constricting the solution space, more quickly converge to a quality solution.

In terms of practical future applications in expert and intelligent systems, the proposed piece-wise regression method can benefit many via automatic extraction of knowledge from databases and generate accurate forecasts. As examples, we have identified the following directions as possible avenues worth investigation in the near future. First, our proposed method can be incorporated into the construction of a decision support expert system that continuously predicts the personalised risk of prisoner with mental illness being released from the jail, aiding clinician for decision making (Constantinou, Freestone, Marsh, Fenton, & Coid, 2015). Other applications that can benefit from our work include intelligent drowsiness monitoring system and stock price prediction. In drowsiness monitoring, the proposed regression model can be built into an intelligent fatigue detection equipment, which records the dynamic physiological signals of drivers or medical staffs and continuously predicts their level of fatigues. A warning will be automatically issued when the model predicts the fatigue level of subjects to be above a prespecified threshold level (Chen, Zhao, Zhang, & zhong Zou, 2015). In financial area, our method can help with construction of an automatic system that forecasts the stock price based on the ever-changing variables quantifying the current performance of a company, including assets, liabilities and income, providing management with data support to make better financial benefits (Ballings, den Poel, Hespeels, & Gryp, 2015). Lastly, the proposed method developed here can also find application in airline industry where managers and decision makers can benefit from a framework powerful of predicting the level of customer satisfaction from various aspects of services, and therefore making it possible for them to carefully allocate resource to maximise customer loyalty (Leong, Hew, Lee, & Ooi, 2015).

Acknowledgments

Funding from the UK Engineering and Physical Sciences Research Council (to LY, SL and LGP through the EPSRC Centre for Innovative Manufacturing in Emergent Macromolecular Therapies), the UK

Leverhulme Trust (to ST and LGP, RPG-2012-686), the European Union (to ST, HEALTH-F2-2011-261366), and the Centre for Process Systems Engineering (CPSE) at Imperial and University College London are gratefully acknowledged.

References

Afantitis, A., Melagraki, G., Sarimveis, H., Koutentis, P. A., Markopoulos, J., & Igglessi-Markopoulou, O. (2006). A novel QSAR model for predicting induction of apoptosis by 4-aryl-4h-chromenes. *Bioorganic and Medicinal Chemistry, 14*(19), 6686–6694.

Alonso, F., Martnez, L., Prez, A., & Valente, J. P. (2012). Cooperation between expert knowledge and data mining discovered knowledge: lessons learned. *Expert Systems with Applications, 39*(8), 7524–7535.

Andrs, J. D., Lorca, P., de Cos Juez, F. J., & Snchez-Lasheras, F. (2011). Bankruptcy forecasting: a hybrid approach using fuzzy c-means clustering and multivariate adaptive regression splines (MARS). *Expert Systems with Applications, 38*(3), 1866–1875.

Bache, K., & Lichman, M. (2013). UCI machine learning repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Bai, Y., Wang, P., Li, C., Xie, J., & Wang, Y. (2014). A multi-scale relevance vector regression approach for daily urban water demand forecasting. *Journal of Hydrology, 517*(0), 236–245.

Ballings, M., den Poel, D. V., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications, 42*(20), 7046–7056.

Balshi, M. S., Mcguire, A. D., Duffy, P., Flannigan, M., Walsh, J., & Melillo, J. (2009). Assessing the response of area burned to changing climate in western boreal north america using a multivariate adaptive regression splines (MARS) approach. *Global Change Biology, 15*(3), 578–600.

Beck, J., Friedrich, D., Brandani, S., Guillas, S., & Fraga, E. (2012). Surrogate based optimisation for design of pressure swing adsorption systems. In *Proceedings of the 22nd European Symposium on Computer Aided Process Engineering, Computer Aided Chemical Enginering, vol. 30, 12* (pp. 1217–1221). Elsevier.

Bermolen, P., & Rossi, D. (2009). Support vector regression for link load prediction. *Computer Networks, 53*(2), 191–201.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research, 13*, 1063–1095.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth.

Brus, D. J., & Heuvelink, G. B. (2007). Optimization of sample patterns for universal kriging of environmental variables. *Geoderma, 138*, 86–95.

Caballero, J. A., & Grossmann, I. E. (2008). An algorithm for the use of surrogate models in modular flowsheet optimization. *AIChE Journal, 54*(10), 2633–2650.

Cavanaugh, K. C., Kellner, J. R., Forde, A. J., Gruner, D. S., Parker, J. D., Rodriguez, W., et al. (2014). Poleward expansion of mangroves is a threshold response to decreased frequency of extreme cold events. *Proceedings of the National Academy of Sciences, 111*(2), 723–727.

Chang, C.-C., & Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 27:1–27:27.

Chen, L., Zhao, Y., Zhang, J., & zhong Zou, J. (2015). Automatic detection of alertness/drowsiness from physiological signals using wavelet-based nonlinear features and machine learning. *Expert Systems with Applications, 42*(21), 7344–7355.

Chen, Q.-L., Wu, K.-J., & He, C.-H. (2014). Thermal conductivity of ionic liquids at atmospheric pressure: Database, analysis, and prediction using a topological index method. *Industrial and Engineering Chemistry Research, 53*(17), 7224–7232.

Cherkassky, V., & Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks, 17*(1), 113–126.

Comrie, A. C. (1997). Comparing neural networks and regression models for ozone forecasting. *Journal of the Air and Waste Management Association, 47*(6), 653–663.

Constantinou, A. C., Freestone, M., Marsh, W., Fenton, N., & Coid, J. (2015). Risk assessment and risk management of violent reoffending among prisoners. *Expert Systems with Applications, 42*(21), 7511–7529.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems, 47*(4), 547–553.

Cozad, A., Sahinidis, N. V., & Miller, D. C. (2014). Learning surrogate models for simulation-based optimization. *AIChE Journal, 60*(6), 2211–2227.

Davis, E., & Ierapetritou, M. (2008). A kriging-based approach to MINLP containing black-box models and noise. *Industrial and Engineering Chemistry Research, 47*(16), 6101–6125.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., et al. (2013). Orange: data mining toolbox in python. *Journal of Machine Learning Research, 14*, 2349–2353.

Dua, V. (2010). A mixed-integer programming approach for optimal configuration of artificial neural networks. *Chemical Engineering Research and Design, 88*(1), 55–60.

Eronen, A., & Klapuri, A. (2010). Music tempo estimation with k-nn regression. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(1), 50–57.

Fanelli, G., Gall, J., & Van Gool, L. (2011). Real time head pose estimation with random regression forests. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR), 2011* (pp. 617–624).

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics, 19*(1), 1–67.

GAMS Development Corporation (2013). *General Algebraic Modeling System (GAMS) Release 24.2.1*. Washington, DC, USA.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters, 31*(14), 2225–2236.

Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling, 160*(3), 249–264.

Ghasemi, J., Saaidpour, S., & Brown, S. D. (2007). Qspr study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. *Journal of Molecular Structure: THEOCHEM, 805*, 27–32.

Greene, M., Rolfson, O., Garellick, G., Gordon, M., & Nemes, S. (2015). Improved statistical analysis of pre- and post-treatment patient-reported outcome measures (proms): the applicability of piecewise linear regression splines. *Quality of Life Research, 24*(3), 567–573.

Gudise, V., & Venayagamoorthy, G. (2003). Comparison of particle swarm optimization and backpropagation as training algorithms for neural networks. In *Proceedings of the 2003 IEEE swarm intelligence symposium, 2003. SIS '03* (pp. 110–117).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explorations Newsletter, 11*(1), 10–18.

Helton, J., & Davis, F. (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering and System Safety, 81*(1), 23–69.

Henao, C. A., & Maravelias, C. T. (2010). Surrogate-based process synthesis. In S. Pierucci, & G. B. Ferraris (Eds.), *Proceedings of the 20th European Symposium on Computer Aided Process Engineering, Computer Aided Chemical Engineering: vol. 28* (pp. 1129–1134). Elsevier.

Henao, C. A., & Maravelias, C. T. (2011). Surrogate-based superstructure optimization framework. *AIChE Journal, 57*(5), 1216–1232.

Hill, T., Marquez, L., O'Connor, M., & Remus, W. (1994). Artificial neural network models for forecasting and decision making. *International Journal of Forecasting, 10*(1), 5–15.

Jekabsons, G. (2015). ARESLab: adaptive regression Splines toolbox for Matlab/Octave. Available at http://www.cs.rtu.lv/jekabsons/.

Khayet, M., Cojocaru, C., & Zakrzewska-Trznadel, G. (2008). Response surface modelling and optimization in pervaporation. *Journal of Membrane Science, 321*(2), 272–283.

Khuri, A. I., & Mukhopadhyay, S. (2010). Response surface methodology. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(2), 128–149.

Kleijnen, J. P. (2009). Kriging metamodeling in simulation: a review. *European Journal of Operational Research, 192*(3), 707–716.

Kleijnen, J. P. C., & Beers, W. C. M. v. (2004). Application-driven sequential designs for simulation experiments: kriging metamodelling. *The Journal of the Operational Research Society, 55*(8), 876–883.

Kone, E. R. S., & Karwan, M. H. (2011). Combining a new data classification technique and regression analysis to predict the cost-to-serve new customers. *Computer and Industrial Engineering, 61*(1), 184–197.

Korhonen, K. T., & Kangas, A. (1997). Application of nearest neighbour regression for generalizing sample tree information. *Scandinavian Journal of Forest Research, 12*(1), 97–101.

Leathwick, J., Elith, J., & Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling, 199*(2), 188–196.

Leong, L.-Y., Hew, T.-S., Lee, V.-H., & Ooi, K.-B. (2015). An semartificial-neural-network analysis of the relationships between servperf, customer satisfaction and loyalty among low-cost and full-service airline. *Expert Systems with Applications, 42*(19), 6620–6634.

Levis, A. A., & Papageorgiou, L. G. (2005). Customer demand forecasting via support vector regression analysis. *Chemical Engineering Research and Design, 83*(8), 1009–1018.

Li, B., Zhang, L., Yan, Q., & Xue, Y. (2014a). Application of piecewise linear regression in the detection of vegetation greenness trends on the tibetan plateau. *International Journal of Remote Sensing, 35*(4), 1526–1539.

Li, S., Feng, L., Benner, P., & Seidel-Morgenstern, A. (2014b). Using surrogate models for efficient optimization of simulated moving bed chromatography. *Computers and Chemical Engineering, 67*(0), 121–132.

Li, Y., Gong, S., & Liddell, H. (2000). Support vector regression and classification based multi-view face detection and recognition. In *Proceedings of the fourth IEEE international conference on automatic face and gesture recognition, 2000.* (pp. 300–305).

Lloyd, C. D., & Atkinson, P. M. (2002). Deriving DSMS from lidar data with kriging. *International Journal of Remote Sensing, 23*(12), 2519–2524.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 14–23.

Lu, C.-J., Lee, T.-S., & Chiu, C.-C. (2009). Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems, 47*(2), 115–125.

Magnani, A., & Boyd, S. (2009). Convex piecewise-linear fitting. *Optimization and Engineering, 10*(1), 1–17.

Malash, G. F., & El-Khaiary, M. I. (2010). Piecewise linear regression: a statistical method for the analysis of experimental adsorption data by the intraparticle-diffusion models. *Chemical Engineering Journal, 163*(3), 256–263.

Matthews, T. J., Steinbauer, M. J., Tzirkalli, E., Triantis, K. A., & Whittaker, R. J. (2014). Thresholds and the species–area relationship: a synthetic analysis of habitat island datasets. *Journal of Biogeography, 41*(5), 1018–1028.

Miller, D. C., Syamlal, M., Mebane, D. S., Storlie, C., Bhattacharyya, D., Sahinidis, N. V., et al. (2014). Carbon capture simulation initiative: a case study in multiscale modeling and new challenges. *Annual Review of Chemical and Biomolecular Engineering, 5*(1), 301–323.

Minjares-Fuentes, R., Femenia, A., Garau, M., Meza-Velzquez, J., Simal, S., & Rossell, C. (2014). Ultrasound-assisted extraction of pectins from grape pomace using citric acid: a response surface methodology approach. *Carbohydrate Polymers, 106*(0), 179–189.

Muggeo, V. M. (2008). Segmented: an r package to fit regression models with broken-line relationships. *R News, 8*(1), 20–25.

Nuchitprasittichai, A., & Cremaschi, S. (2013). An algorithm to determine sample sizes for optimization with artificial neural networks. *AIChE Journal, 59*(3), 805–812.

Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: a review of applications. *Expert Systems with Applications, 36*(1), 2–17.

Palmer, K., & Realff, M. (2002). Metamodeling approach to optimization of steady-state flowsheet simulations: model generation. *Chemical Engineering Research and Design, 80*(7), 760–772.

Pan, J., Kung, P., Bretholt, A., & Lu, J. (2014). Prediction of energys environmental impact using a three-variable time series model. *Expert Systems with Applications, 41*(4, Part 1), 1031–1040.

Papadopoulos, H., Vovk, V., & Gammerman, A. (2011). Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research, 40*(1), 815–840.

Quinlan, J. R. (1992). Learning with continuous classes. In *Proceedings of the australian joint conference on artificial intelligence* (pp. 343–348). World Scientific.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria.

Rafiq, M., Bugmann, G., & Easterbrook, D. (2001). Neural network design for engineering applications. *Computers and Structures, 79*(17), 1541–1552.

Sampson, P. D., Richards, M., Szpiro, A. A., Bergen, S., Sheppard, L., Larson, T. V., et al. A regionalized national universal kriging model using partial least squares regression for estimating annual pm2.5 concentrations in epidemiology. *Atmospheric Environment*, 75(0), 383–392.

Sarimveis, H., Alexandridis, A., Mazarakis, S., & Bafas, G. (2004). A new algorithm for developing dynamic radial basis function neural network models based on genetic algorithms. *Computers and Chemical Engineering, 28*, 209–217.

Scheuber, M. (2010). Potentials and limits of the k-nearest-neighbour method for regionalising sample-based data in forestry. *European Journal of Forest Research, 129*(5), 825–832.

Smola, A., & Schlkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199–222.

Strikholm, B. (2006). Determining the number of breaks in a piecewise linear regression model. Working Paper Series in Economics and Fnance 648 Stockholm School of Economics.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73*(3), 273–282.

Toms, J. D., & Lesperance, M. L. (2003). Piecewise regression: a tool for identifying ecological thresholds. *Ecology, 84*(8), 2034–2041.

Toriello, A., & Vielma, J. P. (2012). Fitting piecewise linear continuous functions. *European Journal of Operational Research, 219*(1), 86–95.

Tsanas, A., & Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings, 49*(0), 560–567.

Venkatesh, K., Ravi, V., Prinzie, A., & den Poel, D. V. (2014). Cash demand forecasting in ATMs by clustering and neural networks. *European Journal of Operational Research, 232*(2), 383–392.

Viana, F. A. C., Simpson, T. W., Balabanov, V., & Toropov, V. (2014). Metamodeling in multidisciplinary design optimization: how far have we really come? *AIAA Journal, 52*(4), 670–690.

Wu, K.-J., Chen, Q.-L., & He, C.-H. (2014). Speed of sound of ionic liquids: Database, estimation, and its application for thermal conductivity prediction. *AIChE Journal, 60*(3), 1120–1131.

Xue, Y., Liu, S., Zhang, L., & Hu, Y. (2013). Integrating fuzzy logic with piecewise linear regression for detecting vegetation greenness change in the yukon river basin, alaska. *International Journal of Remote Sensing, 34*(12), 4242–4263.

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research, 28*(12), 1797–1808.

Zhang, J.-R., Zhang, J., Lok, T.-M., & Lyu, M. R. (2007). A hybrid particle swarm optimization–back-propagation algorithm for feedforward neural network training. *Applied Mathematics and Computation, 185*(2), 1026–1037.

Zhang, Y., & Sahinidis, N. V. (2013). Uncertainty quantification in co2 sequestration using surrogate models from polynomial chaos expansion. *Industrial and Engineering Chemistry Research, 52*(9), 3121–3132.

Zhu, Q., & Lin, H. (2010). Comparing ordinary kriging and regression kriging for soil properties in contrasting landscapes. *Pedosphere, 20*(5), 594–606.